INVERSE PROBLEMS AND MODEL VALIDATION: AN EXAMPLE FROM LATENT VIRUS REACTIVATION

H. T. Banks[a,*], G. M. Kepler[a,†], Hoan K. Nguyen[a,‡] and J. Webster-Cyriaque[b,§]

[a]Center for Research in Scientific Computation
Box 8205
North Carolina State University
Raleigh, N.C. 27695-8205 USA

[b]Department of Dental Ecology, School of Dentistry
University of North Carolina
Chapel Hill, N.C. 27599 USA

August 15, 2006

Abstract

We consider a least squares inverse problem with a model for inducer-mediated reactivation of latent viruses. We illustrate the difficulties associated with the lack of sufficient data (quantity as well as form) in such problems. The modeling and estimation efforts described suggest new experiments as well as needed model extensions.

[*]htbanks@ncsu.edu

[†]gmkepler@ncsu.edu

[‡]hknguyen@ncsu.edu–Invited Lecture, 3rd Intl. Conf. on Inverse Problems, May 29-June 2,2006, Fethiye, Turkey
[§]cyriaquj@dentistry.unc.edu

Report Documentation Page

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE **15 AUG 2006**	2. REPORT TYPE	3. DATES COVERED **00-08-2006 to 00-08-2006**

4. TITLE AND SUBTITLE **Inverse Problems and Model Validation: An Example from Latent Virus Reactivation**	5a. CONTRACT NUMBER
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **North Carolina State University,Center for Research in Scientific Computation,Box 8205,Raleigh,NC,27695-8205**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT **Approved for public release; distribution unlimited**

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**		**17**	

1 Introduction

Mathematical and statistical inverse problem techniques in the context of the biological sciences are becoming increasing prevalent and, consequently, of increasing importance. Classical estimation theory has been primarily developed in a statistical context (based on, e.g., asymptotic distribution, hypothesis testing and Bayesian approaches that involve large data sets) for relatively *simple* models. However, recent advances in both theory and computation suggest that classical inverse problem techniques (regularized least squares, constrained maximum likelihood, etc.) with increasingly *complex* models will play a significant role in near term future biological modeling research efforts. Model development, validation and comparison techniques [8] such as the Kullback-Leibler Information "distance" between models (also called the discrimination information) [16, 17], the Akaike Information Criterion [2, 3], the Takeuchi Information Criterion [24], various Likelihood Ratio Tests [8] and ANOVA type Hypothesis Testing [4, 5] should see significant use with nonlinear dynamical systems models. However, since all of theses statistically–based techniques require the availability of sufficiently rich data sets, they will have limited success and impact in scientific advancement unless inverse problem and estimation "experts" are working in significant, close collaborations with biologists to design experiments in the context of modeling efforts.

In this paper we illustrate some of the inherent difficulties that one can anticipate in trying to develop and validate a reasonably complex dynamical model with only "literature" data (often referred to as "cold data" by inverse problem investigators since it was not collected with model development in mind). We do this in the context of model development of a biologically important system for reactivation of latent virus under the constraint of limited data (no direct observations of any of the seven model compartments, a limited number of longitudinal observations per experiment, observations of only percentages, i.e., ratios, of cell counts). The modeling and least squares efforts lead in this case to specific suggestions for design of new experiments and the type of data needed. In this example, while some parameters can be "estimated" from literature data by considering limiting (as $t \to 0$ or $t \to \infty$) set point values, the longitudinal data available consists of "% viable cells" in terms of ratios of latent plus replicating host cells. This data leads to very difficult least squares or maximum likelihood inverse problems for dynamic parameters as evidenced by relatively large standard errors and to lack of ability to validate viral compartment dynamics in the model. However, as we shall see, the modeling effort leads to suggestions for detailed refined models as well as new experiments.

2 A Mathematical Model

We consider a mathematical model that describes the reactivation of a latent virus by chemical inducers at the cellular level. Here we first give the differential equations that model the dynamics of host cells and viral DNA without deriving the model. Interested readers can find the details and a derivation of the model in [15]. When a latent cell line is growing in a nutritious environment, the cell line is considered to be uninduced

and in the cases considered in [15], can be described by the nonlinear ODEs

$$
\begin{aligned}
\frac{dH_L}{dt} &= (\gamma_L - \alpha_0)H_L \\[2mm]
\frac{dH_R}{dt} &= \alpha_0 H_L - d_I(\bar{V}_I)H_R \\[2mm]
\frac{dH_N}{dt} &= d_L H_L + d_I(\bar{V}_I)H_R - \mu H_N \\[2mm]
\frac{dL}{dt} &= (\gamma_L - \alpha_0)L \\[2mm]
\frac{dR}{dt} &= \alpha_0 L - \left(d_I(\bar{V}_I) - \kappa + b\right)R \\[2mm]
\frac{dV_I}{dt} &= bR - \left(p + d_I(\bar{V}_I)\right)V_I,
\end{aligned}
\tag{1}
$$

and

$$
V_F(t) = V_{F0} + \int_{t_0}^{t} pV_I(u)du.
$$

The compartmental variables and the parameters for this model are defined in Table 1 and Table 2, respectively. Let s denote the concentration level of chemical inducing agents and define $\alpha(s)$ and $\delta_R(s)$ to be the viral reactivation rate and the host cell death rate caused by chemical inducers. Then the host cells and viral dynamics during reactivation (induced) as derived in [15] are governed by the following differential equations

$$
\begin{aligned}
\frac{dH_L}{dt} &= (\gamma_L - \alpha(s))H_L \\[2mm]
\frac{dH_R}{dt} &= \alpha(s)H_L - \left(d_I(\bar{V}_I) + \delta_R(s)\right)H_R \\[2mm]
\frac{dH_N}{dt} &= d_L H_L + \left(d_I(\bar{V}_I) + \delta_R(s)\right)H_R - \mu H_N \\[2mm]
\frac{dL}{dt} &= (\gamma_L - \alpha(s))L \\[2mm]
\frac{dR}{dt} &= \alpha(s)L - \left(d_I(\bar{V}_I) + \delta_R(s) - \kappa + b\right)R \\[2mm]
\frac{dV_I}{dt} &= bR - \left(p + d_I(\bar{V}_I) + \delta_R(s)\right)V_I
\end{aligned}
\tag{2}
$$

and

$$
V_F(t) = V_{F0} + \int_{t_0}^{t} pV_I(u)du.
$$

Compartment	Symbol	Units
Host cells (latent virus only)	H_L	number of cells
Host cells (lytic virus only)	H_R	number of cells
Nonviable host cells	H_N	number of cells
Latent virus	L	DNA copies
Lytic virus	R	DNA copies
Intracellular virus	V_I	DNA copies
Free virus	V_F	number of virions

Table 1: ODE Model Compartments.

Parameter	Symbol	Units
Net growth of latent host cells	γ_L	hr^{-1}
Nonviable cell degradation	μ	hr^{-1}
Natural death of latent host cells	d_L	hr^{-1}
Spontaneous reactivation of latent host cells	α_0	hr^{-1}
Cell death due to viral lysis	c	hr^{-1}
Synthesis of viral DNA	κ	hr^{-1}
Sequestration of viral DNA for encapsulation	b	hr^{-1}
Packaging and secretion of virions	p	hr^{-1}
Viral DNA per lytic host cell	n_T	-

Table 2: Uninduced Parameter Description.

3 Parameter Estimation

We first estimate many of the uninduced model parameters from experimental observations cited in the literature. In the uninduced cell cultures, experimentalists normally observe the fraction a_s of lytic host cells, the fraction N_r of nonviable host cells, host cell doubling time D_p and average copies n_T of viral DNA per cell. These values can be approximated using values from the experimental literature [10, 14, 18, 19, 20, 21, 25, 27] and they mathematically correspond to

$$a_s = \left(\frac{H_R}{H_L + H_R + H_N} \right)_{t \to \infty} \qquad N_r = \left(\frac{H_N}{H_L + H_R + H_N} \right)_{t \to \infty}$$

$$2 = \frac{(H_L + H_R + H_N)_{t=D_p}}{(H_L + H_R + H_N)_{t=0}} \qquad n_T = \left(\frac{L + R + V_I}{H_L + H_R + H_N} \right)_{t \to \infty}. \tag{3}$$

It follows that the uninduced parameters are related by

$$d_L = \frac{\ln(2)/D_p + \mu N_r}{1 - a_s - N_r} - \gamma_L$$

$$c = \frac{\ln(2)}{a_s V_{IA} D_p}(N_r - 1) + \frac{\gamma_L}{a_s V_{IA}}(1 - a_s - N_r)$$

$$\alpha_0 = \gamma_L - \frac{\ln(2)}{D_p}$$

$$p = \frac{R_A}{T V_{IA}} - \frac{(1 - a_s - N_r)}{a_s}\left(\gamma_L - \frac{\ln(2)}{D_p} \right) \tag{4}$$

$$\kappa = \left(\frac{\gamma_L - \ln(2)/D_p}{a_s R_A} \right)(R_A - N_r R_A - n_T + a_s V_{IA}) + b$$

$$n = \frac{n_T - a_s(R_A + V_{IA})}{1 - a_s - N_r},$$

where $n = L/H_L$ denotes the average number of latent viral DNA copies per latently infected host cell which is assumed to be constant for the model derivation in [15]. Here R_A and V_{IA} denote the average viral replicating DNA, respectively, and viral intracellular DNA and are defined by

$$V_{IA} = \left(\frac{V_I}{H_R} \right)_{t \to \infty} \qquad R_A = \left(\frac{R}{H_R} \right)_{t \to \infty}. \tag{5}$$

Although the values of R_A and V_{IA} are the only two values in the uninduced model parameters that cannot be obtained from the literature; they are chosen for our simulations so that (4) holds and all the parameters involved are positive.

We next formulate an inverse problem for the induced model to obtain the viral reactivation rate $\alpha(s)$ and the host cell death rate $\delta_R(s)$ by the chemical inducers using cell viability data. We use two independent sets of experimental data from the literature [26, 27] wherein the data are given in percentage of viable cells at different inducer (butyrate) concentration and are collected every 24 hours over a maximum period of five days.

We considered the special case with parametric functional forms $\delta_R(s) = \delta_c s$, $\alpha(s) = \alpha_c s + \alpha_0$ and let $q = (\delta_c, \alpha_c)$ and $x = (H_L, H_R, H_N, L, R, V_I, V_F)^T$. Then the differential equations in the model for the induced case (2) can be written in a general form

$$\dot{x} = g(t, x, s, q) \qquad (6)$$
$$x(0) = x_0,$$

where $g : \mathbb{R}_+ \times \mathbb{R}^r \times \mathbb{R}_+ \times \mathbb{R}^m \to \mathbb{R}^r$ for $r = 7$, $m = 2$, and $x_0 = (H_{L0}, H_{R0}, H_{N0}, L_0, R_0, V_{I0}, V_{F0})^T$. To correspond with the experimental data given in percentage of viable cells, we define the outputs of the model

$$f(t, s, q) = \left[\frac{H_{total}(t, s, q) - H_N(t, s, q)}{H_{total}(t, s, q)} \right], \qquad t, s \geq 0,$$

where $H_{total} = H_L + H_R + H_N$. In each least squares parameter fit to the data, the data is longitudinal (taken at t_k) and across several levels s_i of inducer. This is indexed by $\tau_j = (t_k, s_i)$ for $k = 1, \ldots, K$, $i = 1, \ldots, I$, and observations y_j for the model values $f_j(q) = f(t_k, s_i, q)$, $j = 1, \ldots, N = KI$. Then we construct the ordinary least square (OLS) inverse problem by seeking $\hat{q}$ that minimize the cost criterion

$$J(q) = \sum_{j=1}^{N} |y_j - f_j(q)|^2, \qquad (7)$$

where $\{y_j\}$ denotes the experimental data.

Once the optimal $\hat{q}$ are found using the Nelder-Mead algorithm, standard errors and confidence intervals are computed by using the asymptotic theory which we proceed to outline. Assume N scalar longitudinal/inducer level observations (time/inducer series of numbers or ratios of numbers of cells as described below) are represented by the statistical model

$$Y_j \equiv f_j(q_0) + \epsilon_j, \quad j = 1, 2, \ldots N, \qquad (8)$$

where $f_j(q_0)$ is the model for the observations in terms of the state variables and $q_0 \in \mathbb{R}^m$ is a "set" of theoretical "true" parameter values (assumed to exist in a standard statistical approach). Assume further that the errors ϵ_j, $j = 1, 2, \ldots, N$ in the statistical model of the observation or measurement process (8) are independent identically distributed (*i.i.d.*) random variables with mean $E[\epsilon_j] = 0$ and constant variance $var[\epsilon_j] = \sigma_0^2$ where of course σ_0^2 is unknown. The constant variance assumption can be validated by use of standard residual plots with the data used in our inverse problems. It follows that the observations Y_j are *i.i.d.* with mean $E[Y_j] = f_j(q_0)$ and variance $var[Y_j] = \sigma_0^2$.

Using the data $\{y_j\}$ for the observation process $\{Y_j\}$ with the model, $J(q)$ is optimized by finding the OLS estimator $\hat{q}$ in (7). Note that the estimator $\hat{q}_{OLS}$ is also a random variable with a distribution called the *sampling distribution* because Y_j is a random variable. Knowledge of this sampling distribution provides uncertainty information (e.g., standard errors) for the numerical values of $\hat{q}$ obtained using a specific data set $\{y_j\}$ (i.e., a realization of $\{Y_j\}$) when minimizing $J(q)$.

Under reasonable assumptions on smoothness and regularity (the smoothness requirements for model solutions are readily verified using continuous dependence results for ordinary differential equations in our example; the regularity requirements involve, among others, conditions on how the observations are taken as sample size increases, i.e., as $N \to \infty$), the standard nonlinear regression approximation theory ([11], [12], [13], and Chapter 12 of [22]) for asymptotic (as $N \to \infty$) distributions can be invoked. This theory yields that the sampling distribution for the estimate $\hat{q}(Y) = \hat{q}_{OLS}(Y)$, where $Y = \{Y_j\}_{j=1}^N$, is approximately a m-multivariate Gaussian with mean $E[\hat{q}(Y)] \approx q_0$ and covariance matrix $cov[\hat{q}(Y)] \approx \Sigma_0 = \sigma_0^2 [\chi^T(q_0)\chi(q_0)]^{-1}$. Here $\chi(\hat{q}) = F_q(q)$ is the $N \times m$ sensitivity matrix with elements

$$\chi_{jk}(q) = \frac{\partial f_j(q)}{\partial q_k} \qquad \text{and} \qquad F_q(q) \equiv (f_{1q}(q), \ldots, f_{Nq}(q))^T.$$

That is, for N large, the sampling distribution approximately satisfies

$$\hat{q}_{OLS}(Y) \sim \mathcal{N}_m(q_0, \sigma_0^2[\chi^T(q_0)\chi(q_0)]^{-1}) := \mathcal{N}_m(q_0, \Sigma_0). \tag{9}$$

The elements of the matrix $\chi = (\chi_{jk})$ can be estimated using the forward difference

$$\chi_{jk}(q) = \frac{\partial f_j(q)}{\partial q_k} \approx \frac{f_j(q + h_k) - f_j(q)}{|h_k|},$$

where h_k is an m-vector with nonzero entry in only the kth component, or using sensitivity equations (see [7] and the references therein). Here we chose the sensitivity equation approach and since q_0, σ_0 are not known, we approximate them in $\Sigma_0 = \sigma_0^2[\chi^T(q_0)\chi(q_0)]^{-1}$. Following standard practice, Σ_0 is approximated by

$$\Sigma_0 \approx \Sigma(\hat{q}) = \hat{\sigma}^2[\chi^T(\hat{q})\chi(\hat{q})]^{-1}$$

where $\hat{q}$ is the parameter estimate obtained from minimizing (7) and $\chi(\hat{q}) = \dfrac{\partial F}{\partial q}(\hat{q})$. From the outputs defined in (6), it suffices to compute the sensitivities $\dfrac{\partial x}{\partial q}$ by solving the $r \times m$ matrix *linear variational differential equation* (called the *the sensitivity equation* in the applied mathematics and engineering literature)

$$\frac{d}{dt}\left(\frac{\partial x}{\partial q}\right) = \frac{\partial g}{\partial x}\frac{\partial x}{\partial q} + \frac{\partial g}{\partial q}. \tag{10}$$

The matrix coefficient and the forcing function in this equation are evaluated along solutions of the system equation (6). The approximation $\hat{\sigma}^2$ to σ_0^2 is given by

$$\sigma_0^2 \approx \hat{\sigma}^2 = \frac{1}{N - m}\sum_{j=1}^{N}|y_j - f_j(\hat{q})|^2.$$

Standard errors to be used in confidence interval calculations are thus given by $SE_k(\hat{q}) = \sqrt{\Sigma_{kk}(\hat{q})}$, $k = 1, 2, \ldots, m$ (see [9]).

Finally, in order to compute the confidence intervals (at the $100(1 - c)\%$ level) for the estimated parameters δ_c and α_c, we define the confidence level parameters associated with the estimated parameters so that

$$P(\hat{q}_k - t_{c/2}SE_k(\hat{q}) < q_k < \hat{q}_k + t_{c/2}SE_k(\hat{q})) = 1 - c, \tag{11}$$

where $c \in [0, 1]$, and $t_{c/2} \in \mathbb{R}_+$. For a given c value (small, say $c = .05$ for 95% confidence intervals), the critical value $t_{c/2}$ is computed from the Student's t distribution t^{N-m} with $N - m$ degrees of freedom since for each of the data sets available to us we have $N < 30$. The value of $t_{c/2}$ is determined by $P(T \geq t_{c/2}) = c/2$ where $T \sim t^{N-m}$.

4   Numerical Results

In the uninduced experiments, Zoeteweij's group observed 7.7% nonviable cells while Yu, *et al.*, reported 16% nonviable cells in their data. Therefore, we let $N_r = 0.077$ and 0.16, $n_T = 69$ and 68 to correspond with the data of Zoeteweij, *et al.*, and Yu, *et al.*, respectively. Furthermore, the fraction (percentage) of spontaneous

lytic host cells a_s and host cell doubling time D_p are chosen to be within the ranges reported in the literature, i.e., $a_s = 0.02$ and $D_p = 24$ hours. In Figure 1, we depict the percentage of nonviable cells and spontaneously reactivated host cells from the simulation of the uninduced model case (1) for the Zoeteweij, *et al.*, and the Yu, *et al.*, experimental data. The initial condition for all compartments are zero except for $H_{L0} = 1.0 \times 10^6$ and $L_0 = 1.0 \times 10^7$. The values of the parameters used in the simulations presented here are tabulated in Table 3. From Figure 1, we observe that the nonviable cell and spontaneously reactivated cell percentages asymptotically reach the specified "equilibrium" values by 1000 hours where we define "equilibrium" to be the constants in (3) and (5). Even though the cell culture is growing exponentially, the characteristics of the cell culture eventually become those constants in (3) and (5). The simulations of the average number of lytic and intracellular DNA per lytically infected host cell, R/H_R and V_I/H_R, respectively are shown in Figure 2. Similarly, we see both R/H_R and V_I/H_R converge to the specified equilibrium values of $R_A = 2500$ and $V_{IA} = 500$, respectively by 1000 hours.

The equilibrated simulations and the estimated parameters in Table 3 of the uninduced model case are assumed to be the properties of the uninduced (control) cell cultures that are used in the reactivation experiments. In the induced model case, let C_0 be the initial number of cells; then the initial conditions are set to be $H_L(0) = (1 - N_r - a_s)C_0$, $H_R(0) = a_sC_0$, $H_N(0) = N_rC_0$, $L(0) = nH_L$, $R(0) = R_AH_R$, $V_I(0) = V_{IA}H_R$, and $V_F(0) = 0$. As mentioned above, the functional form of the induced lytic cell death rate $\delta_R(s)$ and reactivation rate $\alpha(s)$ are assumed to be affine functions of the form $\delta_R(s) = \delta_c s$ and $\alpha(s) = \alpha_c s + \alpha_0$ as a first approximation to the "true" forms which are not known. The optimal induced parameters $\hat{q} = [\hat{\delta}_c, \hat{\alpha}_c]$ are obtained by fitting the two sets of experimental data from Zoeteweij, *et al.*, [27] and Yu, *et al.*, [26] independently in minimizing (7). The optimized induced model parameters are tabulated in Table 4 along with the standard errors and confidence intervals which are calculated from the mathematical and statistical method presented in Section 3. We hasten to caution that here we had at most $N = 16$ observations when estimating the two parameters (δ_c, α_c) and thus the following remarks result from using asymptotic distributions when it may be unwarranted. Nonetheless, from Table 4, we see that the values of the estimated parameters of both groups are within an order of magnitude of each other. In addition, for both experimental data sets, the standard errors of the estimated reactivation rate constants $\hat{\alpha}_c$ are at least one order of magnitude less than the parameter values while the standard errors of the induced death rate constants $\hat{\delta}_c$ have the same order of magnitude as the parameter values. Thus, we might argue for more confidence in the values obtained for $\hat{\alpha}_c$ than values obtained for $\hat{\delta}_c$. (But this could also be a result of using asymptotic analysis when insufficient data has been used.) The estimated induced parameters $\hat{q} = [\hat{\delta}_c, \hat{\alpha}_c]$ are then used to generate the viable cell percentage with (2) and then plotted against the experimental data from Zoeteweij, *et al.*, and Yu, *et al.*, in Figure 3. It can be seen from Figure 3 that the induced simulations qualitatively agree with both sets of experimental data.

The number of free virions (V_F) simulated from the induced equation (2) with optimized parameters $\hat{q}$ are plotted in Figure 4. From the figure, we observe that there is a three to four-fold increase in free virions produced at 0.3mM butyrate level compared to 3mM butyrate concentration level. This observation in the simulations agree qualitatively with the experiments from Yu, *et al.* In [26], Yu and his group report that with high butyrate concentrations (1.5 and 3 mM), there is a great increase in lytic activity but also a significant increase in cell death. Therefore, very few free virions are produced in the experiments due to massive cell death before the completion of the lytic program. However, with a lower concentration level of butyrate (≤ 0.3 mM) they observe much less cell death and a significant secretion of free virions. The optimal butyrate dosage threshold that maximizes viral production is numerically computed from equation (2) and depicted in Figure 5 for the two data sets. From Figure 5 we interpret that the butyrate concentrations below the optimal threshold suggest that the reactivation activity of latent virus by the inducers is not maximized. On the other hand, the butyrate concentrations above the threshold imply that the concentration of the inducers is too toxic and the lytically replicating cells are killed before virus is produced.
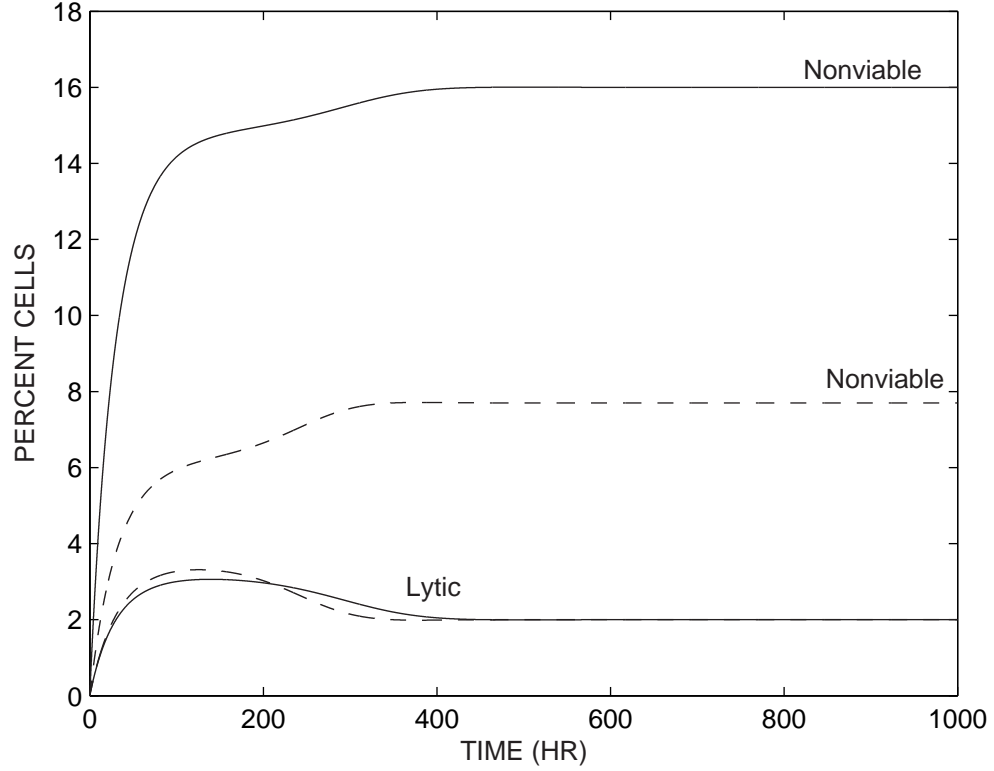
8

Figure 1: Uninduced simulations of spontaneously reactivated cell $(100 \times H_R/(H_{\text{total}}))$ and nonviable cell percentage $(100 \times H_N/(H_{\text{total}}))$ for the data of Zoeteweij (dashed lines), *et al.*, and Yu (solid lines), *et al.*

Parameter	Symbol	Zoeteweij, *et al.*, data	Yu, *et al.*, data	Units
Net growth of latent host cells	γ_L	3.00×10^{-2}	3.00×10^{-2}	hr^{-1}
Nonviable cell degradation	μ	1.00×10^{-5}	1.00×10^{-5}	hr^{-1}
Natural death of latent host cells	d_L	1.98×10^{-3}	5.22×10^{-3}	hr^{-1}
Spontaneous reactivation of latent host cells	α_0	1.12×10^{-3}	1.12×10^{-3}	hr^{-1}
Cell death due to viral lysis	c	4.33×10^{-5}	3.40×10^{-5}	hr^{-1}
Synthesis of viral DNA	κ	7.11×10^{-2}	6.65×10^{-2}	hr^{-1}
Sequestration of viral DNA for encapsulation	q	2.08×10^{-2}	2.08×10^{-2}	hr^{-1}
Packaging and secretion of virions	p	5.36×10^{-2}	5.83×10^{-2}	hr^{-1}
Viral DNA per lytic host cell	n_T	69	68	-
Induced reactivation	$\hat{\alpha}_c$	5.51×10^{-1}	1.40×10^{-1}	hr^{-1}
Induced death	$\hat{\delta}_c$	5.13×10^{-3}	6.76×10^{-3}	hr^{-1}

Table 3: Parameters from the uninduced model (1) are calculated from (4) with constants $a_s = 0.02$, $N_r = 0.077$ or 0.16, $D_p = 24$ hr, $n = 10$, $V_{IA} = 500$, and $R_A = 2500$. Parameters from the induced model (2) are obtained from fits to experimental data.
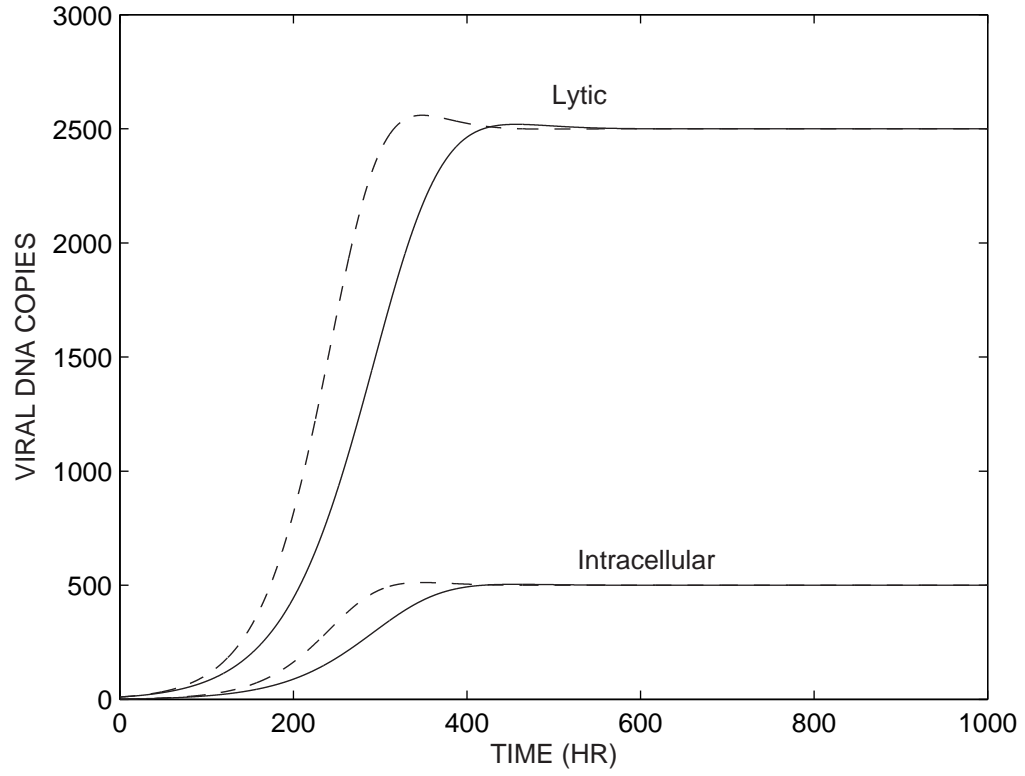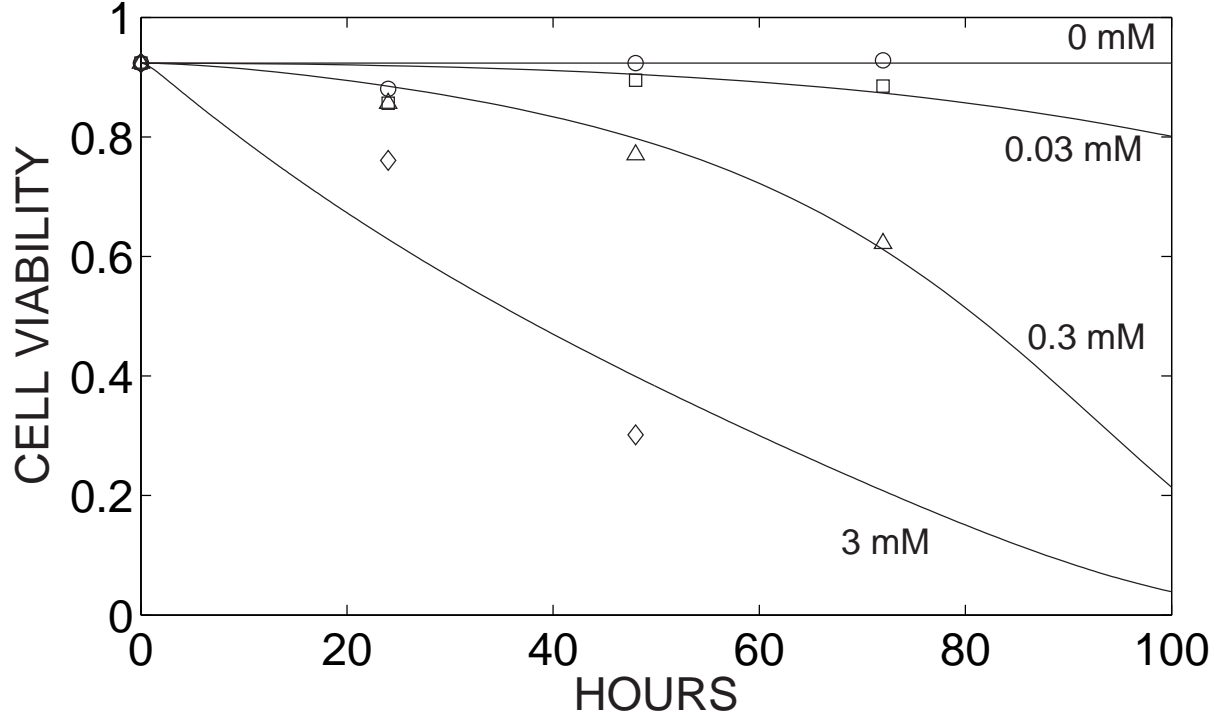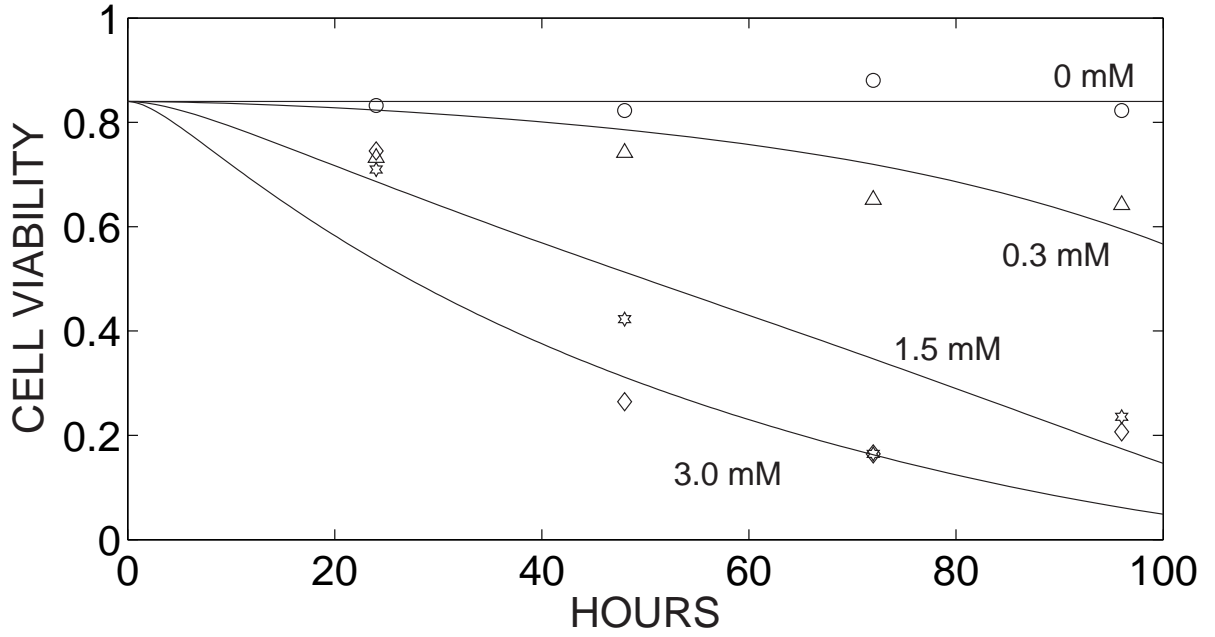
9

Figure 2: Uninduced simulations of the average number of lytic viral DNA copies R/H_R and the average number of intracellular viral DNA copies V_I/H_R per lytically infected host cell for the data of Zoeteweij (dashed lines), *et al.*, and Yu (solid lines), *et al.*
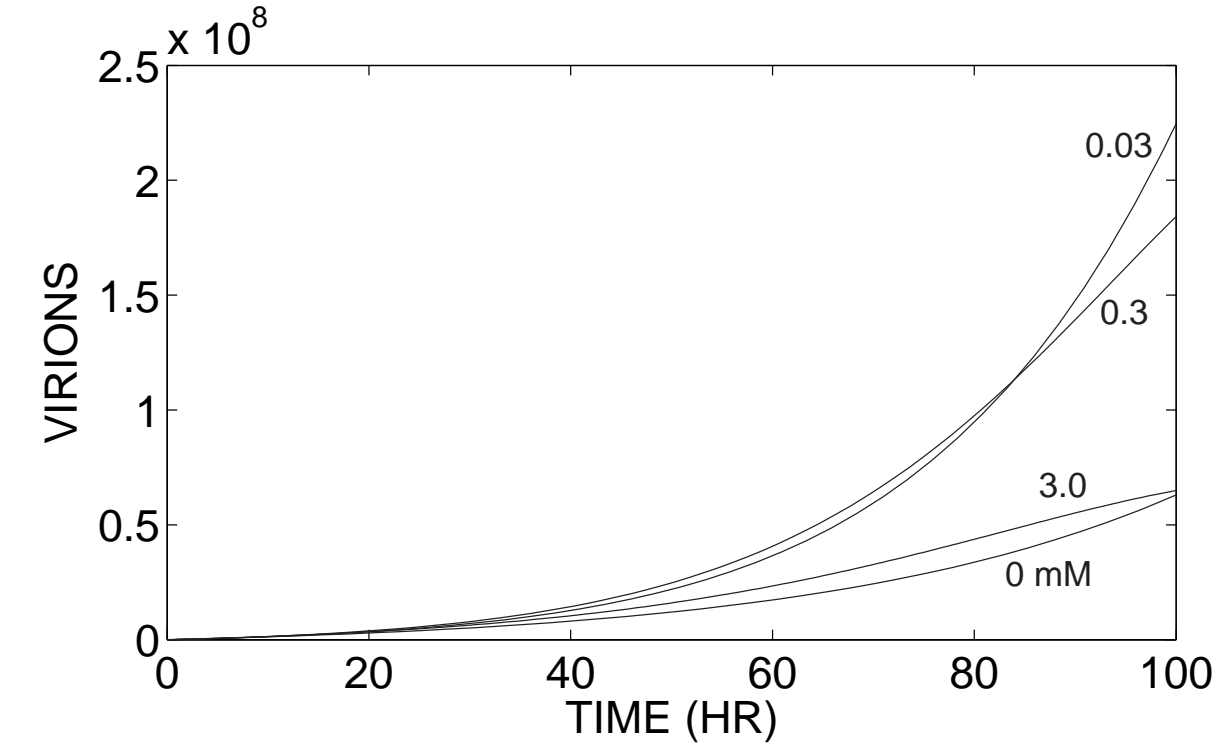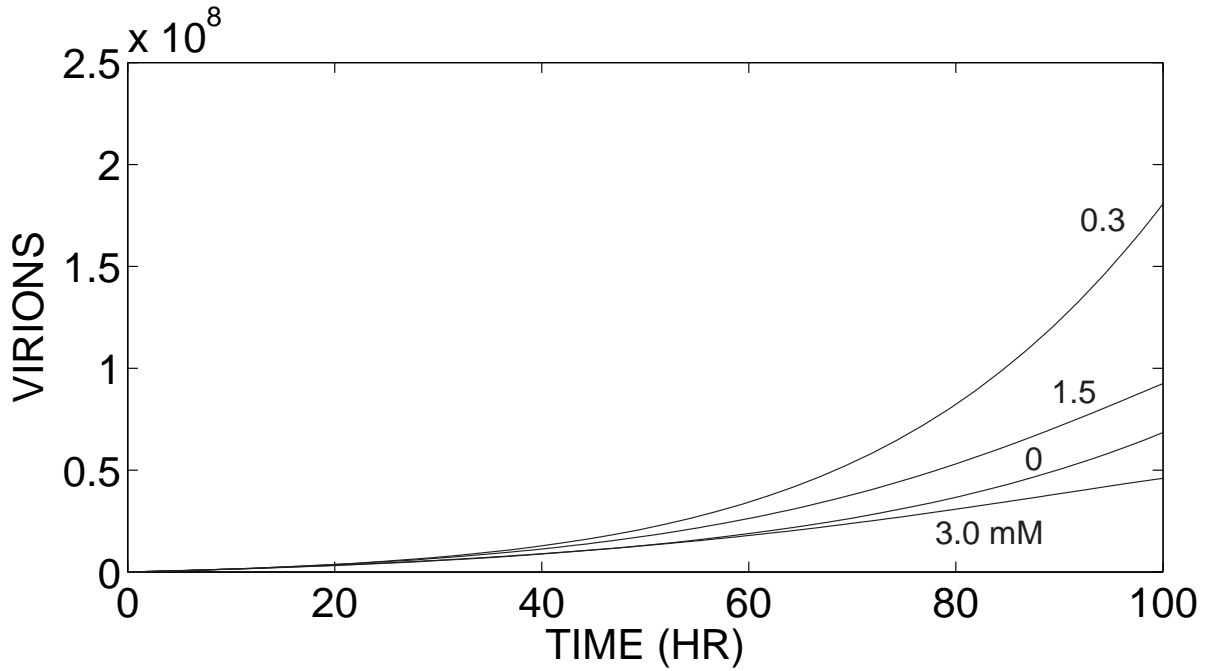
Figure 3: Cell viability experimental measurements (symbols) and induced model (2) simulations (solid lines) using fitted parameters for linear functions $\alpha(s)$ and $\delta_R(s)$: (a) Zoeteweij, *et al.*, circles (0 mM), squares (0.03 mM), triangles (0.3 mM), and diamonds (3 mM), $\hat{\alpha}_c = 0.551$, $\hat{\delta}_c = 5.13 \times 10^{-3}$ and (b) Yu, *et al.*, circles (0 mM), triangles (0.3 mM), stars (1.5 mM), and diamonds (3 mM), $\hat{\alpha}_c = 0.140$, $\hat{\delta}_c = 6.76 \times 10^{-3}$.

Figure 4: Induced model (2) simulations of free virions using optimized parameters for linear functions $\alpha(s)$ and $\delta_R(s)$ for experimental data from (a) Zoeteweij, *et al.*, and (b) Yu, *et al.*
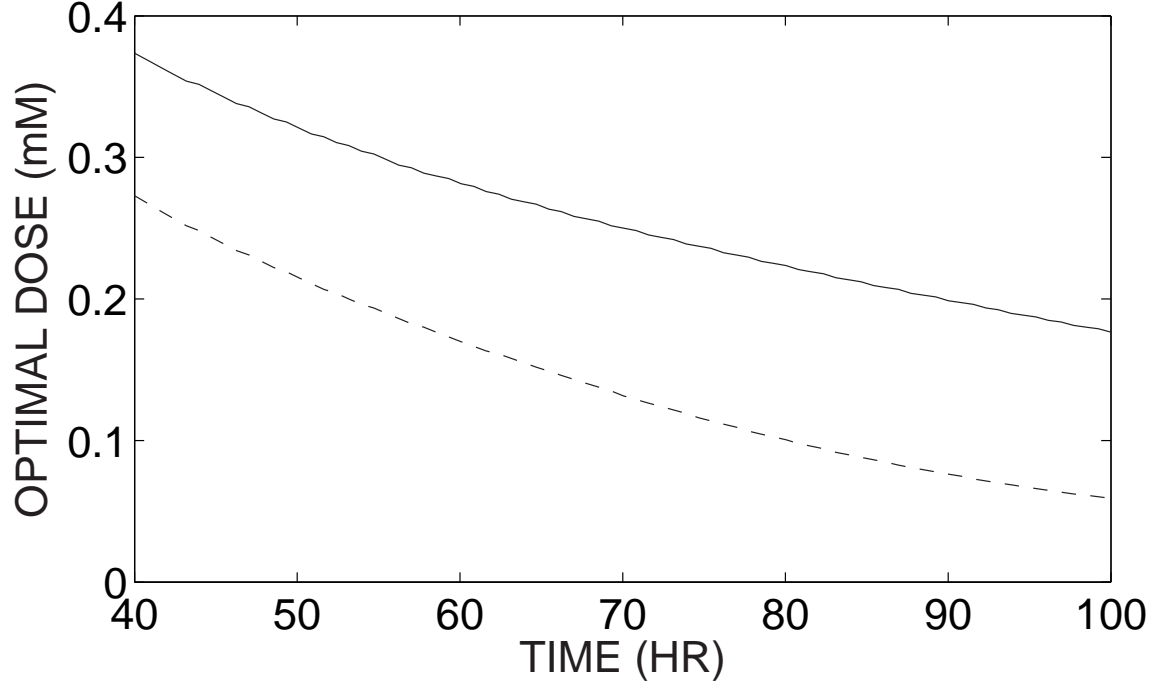
Figure 5: The optimal butyrate concentration to maximize the quantity of virions produced as a function of the elapsed induction time for data from Zoeteweij (dashed line), *et al.*, and Yu (solid line), *et al.*

Data	Parameter	Estimated Value	Standard Error	Confidence Interval
Zoeteweij, *et al.*	$\hat{\alpha}_c$	5.51×10^{-1}	1.03×10^{-2}	$[5.29 \times 10^{-1}, 5.73 \times 10^{-1}]$
	$\hat{\delta}_c$	5.13×10^{-3}	3.25×10^{-3}	$[-1.83 \times 10^{-3}, 1.21 \times 10^{-2}]$
Yu, *et al.*	$\hat{\alpha}_c$	1.40×10^{-1}	7.84×10^{-3}	$[1.23 \times 10^{-1}, 1.57 \times 10^{-1}]$
	$\hat{\delta}_c$	6.76×10^{-3}	2.88×10^{-3}	$[5.30 \times 10^{-4}, 1.30 \times 10^{-2}]$

Table 4: Estimated parameter values, standard errors, and confidence intervals

5   Discussion and Concluding Remarks

After carrying out the inverse problem with this preliminary model, numerical simulations using the resulting estimated parameters provide good qualitative fits of cell viability for two independent data sets of KSHV being induced by butyrate. However, our findings also strongly motivate the *need for more longitudinal data* to support model development and validation. With the cell viability data now available, we can only directly validate the host cells dynamics and *indirectly* justify the dynamics of the viral DNA in this model. Specifically, quantitative experimental *measurements of cell-associated DNA ($L + R + V_I$) and free virions (V_F) are needed* to evaluate model predictions for the viral compartments. We note that the experimental data we require are not the typical data collected in experiments. In other words, typical experimental data are relatively qualitative data recorded at one time point while quantitative longitudinal data are needed to verify the host cell and viral dynamics presented in this model. Indeed, in almost all efforts with *dynamical* models, longitudinal data is essential. These time series data would also permit determination (i.e., estimation) of the free parameters γ_L and μ, as well as the unknown constants R_A and V_{IA}. We also note that R_A and V_{IA} parameter sensitivity tests reveal that the optimal parameter values $\hat{\delta}_c$ and $\hat{\alpha}_c$ are relatively insensitive to variations in R_A or V_{IA}, since varying R_A or V_{IA} by $\pm$ 5% produced 3% or less variation in the optimized parameter values from the corresponding inverse problems.

In addition to the fits of the linear form for $\alpha(s)$ and $\delta_R(s)$, we also fitted the model in OLS problems where $\alpha(s)$ and $\delta_R(s)$ are allowed to be of Michaelis-Menton form and/or sigmoid shape functions. In Figure 3 from the previous section, we see that the fits to cell viability are reasonable with the simple linear functions for $\alpha(s)$ and $\delta_R(s)$. We also found that the fits of the induced equations to cell viability data were relatively insensitive to more complicated functional forms of $\alpha(s)$ and $\delta_R(s)$ (the results are not shown here). In the future, instead of estimating $\alpha(s)$ and $\delta_R(s)$ with some *a priori* parameterizations, we could estimate shape of the functional form using piece-wise linear splines or other approximations as has been successfully done in other problems with temporally varying parameters and dynamic coefficients [1, 5]. However, to validate the improvement provided by such models with sophisticated model comparison techniques, richer data sets are again needed.

With this preliminary mathematical model as an initial study of the reactivating mechanism of latent viruses by chemical inducers, there are rather obvious modifications we would like to pursue in developing future generations of the model. First, instead of assuming a constant (net) growth rate for the host cells, a nonconstant assumption on the host cells (net) growth rate should be added especially after 48 hours. Also, the initial model presented here is based on the assumption that the host cells are of "all or nothing" type. That is, a given host cell either has all latent viral DNA (H_L) or all lytic replicating DNA (H_R) in the nucleus. A more realistic situation can be illustrated by superimposing a probability distribution on the parameters to better approximate mixed conditions where a host cell may contain both latent and lytic virus in varying levels. Such a modeling technique was successfully used in [6] for cellular level HIV models to account for variable length (with uncertainty) pathways. In models of this type the state variables are the *expected values* of concentrations (or of numbers of cells) resulting in delay differential equation models embodying uncertainty through the explicit dependence of dynamics on probability distributions. Once again, richer data sets are essential to establish validity for such models.

Finally, instead of a single viral compartment R to quantify copies of viral DNA in the lytic program, we suggest that one might modify the model to describe Immediate Early, Early, and Late gene expression

(RNA), represented by compartments R_1, R_2, and R_3 for an induced model in the form

$$
\begin{aligned}
\frac{dH_L}{dt} &= \left(\gamma_L - \alpha(s) - \delta_L(s)\right) H_L + \rho H_R \\[2ex]
\frac{dH_R}{dt} &= \left(\gamma_R - \delta_R(s) - d_I(\bar{V}_I) - \rho\right) H_R + \alpha(s) H_L \\[2ex]
\frac{dH_N}{dt} &= \left(d_L + \delta_L(s)\right) H_L + \left(d_R + \delta_R(s) + d_I(\bar{V}_I)\right) H_R - \mu H_N \\[2ex]
\frac{dL}{dt} &= \left(\gamma_L - \alpha(s) - \delta_L(s)\right) L + \rho(R_1 + R_2 + R_3) \\[2ex]
\frac{dR_1}{dt} &= \left(\gamma_R - b_1 - \delta_R(s) - d_I(\bar{V}_I) - \rho\right) R_1 + \alpha(s) L \\[2ex]
\frac{dR_2}{dt} &= \left(\kappa + \gamma_R - b_2 - \delta_R(s) - d_I(\bar{V}_I) - \rho\right) R_2 + b_1 R_1 \\[2ex]
\frac{dR_3}{dt} &= \left(\gamma_R - b_3 - \delta_R(s) - d_I(\bar{V}_I) - \rho\right) R_3 + b_2 R_2 \\[2ex]
\frac{dV_I}{dt} &= b_3 R_3 - \left(p + d_R + d_I(\bar{V}_I) + \delta_R(s)\right) V_I
\end{aligned}
\tag{12}
$$

and $V_F(t) = V_{F0} + \int_{t_0}^{t} p V_I(u) du$. The three parameters b_1, b_2, and b_3 represent the rate at which viral DNA moves from one stage of the lytic program to the next. These parameters can be estimated as $1/T_1$, $1/T_2$, and $1/T_3$, respectively, where T_1, T_2, and T_3 are the approximate times for each stage of the lytic program. Again, one would expect variability in these times across cell populations, suggesting a desired introduction of probability distributions to be estimated instead of the times themselves. Corresponding parameters in this proposed model and model (2) would not, of course, necessarily represent the same quantities.

By having model compartments that quantify RNA production or promoter activity from genes representative of each stage of the lytic cycle, one could hope to predict viral reactivation in more detail and compare to experimental gene expression data. For example, ORF50, vIL6, K8.1 [23] could be representative of the Immediate Early, Early, and Late stages, respectively. A single compartment L can represent latent gene expression, primarily ORF73 expression [23].

There may be underlying biological delays, due to the ordered cascade of gene expression that makes up the lytic program, that are not captured with the model (2). A model such as (12) in which we rewrite the single R compartment as three compartments R_1, R_2, and R_3 representing the Immediate Early, Early, and Late phases of the lytic program might be expected to more closely capture the biological delays (or the associated probability distributions) inherent in the lytic program of the system. To enable development and validation of such models, experiments involving multi-compartment longitudinal observations will be required.

In summary, the results reported on here illustrate well a common finding in initial modeling efforts for biological systems: the need for more *longitudinal data* and the need for *different types of observations* than

are prevalent in current experiments (such as those involving reactivation of latent viruses). Regarding the *initial modeling aspects* of this project, we recall that most modeling efforts (including this one) are steps in an *iterative process* in which one takes experimental observations and forms statistical and mathematical models. These models, when used in inverse problems, suggest inadequacies in both modeling and the experimental data collected. This leads to model modification and extension as well as new experiments to collect data necessary to validate the new model.

Acknowledgements

References

[1] B.M. Adams, *Non-parametric Parameter Estimation and Clinical Data Fitting With a Model of HIV Infection,* Ph.D. Thesis, Center for Research in Scientific Computation, Mathematics Department, North Carolina State University, 2005.

[2] H. Akaike, Information theory as an extension of the maximum likelihood principle, in *Second International Symposium on Information Theory*, B.N. Petrov and F. Csaki, eds., Akademiai Kiado, Budapest, 1973, p. 267–281.

[3] H. Akaike, A new look at the statistical model identification, *IEEE Transactions on Automatic Control*, AC **19** (1974), 716–723.

[4] H.T. Banks and B. Fitzpatrick, Statistical methods for model comparison in parameter estimation problems for distributed systems, *J. Math. Biology*, **28** (1990), 501–527.

[5] H.T. Banks and K. Kunisch, *Estimation Techniques for Distributed Parameter Systems*, Birkhaüser, Boston, 1989.

[6] H.T. Banks, D.M. Bortz and S.E. Holte, Incorporation of variability into the mathematical modeling of viral delays in HIV infection dynamics, *Mathematical Biosciences*, **183** (2003), 63–91.

[7] H.T. Banks and H.K. Nguyen, Sensitivity of dynamical system to Banach space parameters, *CRSC-TR05-13*, NCSU, February, 2005; *J. Math Anal. Appl.*, to appear.

[8] K.P. Burnham and D.R. Anderson, *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, Springer-Verlag, New York, 2002.

[9] G. Casella and R.L Berger, *Statistical Inference*, Duxbury, CA, 2002.

[10] S.R. Chan, C. Bloomer and B. Chandran, Identification and characterization of human herpesvirus-8 lytic cycle-associated ORF59 protein and the encoding cDNA by monoclonal antibody, *Virology* **240** (1998), 118–126.

[11] M. Davidian and D.M. Giltinan, *Nonlinear Models for Repeated Measurement Data*, Chapman and Hall, London, 1995.

[12] A.R. Gallant, *Nonlinear Statistical Models*, John Wiley & Sons, Inc., NY, 1987.

[13] R.I. Jennrich, Asymptotic properties of non-linear least squares estimators, *Ann. Math. Statist.*, **40** (1969), 633–643.

[14] C.M. Klass, L.T. Krug, V.P. Pozharskaya and M.K Offermann, The targeting of primary effusion lymphoma cells for apoptosi by inducing lytic replication of human herpesvirus 8 while blocking virus production, *Blood* **105** (2005), 4028–4034.

[15] G.M. Kepler, H.K. Nguyen, J. Webster-Cyriaque and H.T. Banks, A dynamic model for induced reactivation of latent virus, *CRSC-TR05-44*, December, 2005; *J. Theoretical Biology*, to appear.

[16] S. Kullback, The Kullback-Leibler distance, *The American Statistician*, **41** (1987), 340–341.

[17] S. Kullback and R.A. Leibler, On information and sufficiency, *Annals of Mathematical Statistics*, **22** (1951), 79–86.

[18] F. Lallemand, N. Desire, W. Rozenbaum, J.C. Nicolas and V. Marechal, Quantitative analysis of human herpesvirus 8 viral load using real-time PCR assay, *Journal of Clinical Microbiology*, **38** (2000), 1404–1408.

[19] M. Lu, J. Suen, C. Frias, R. Pfeiffer, M.H. Tsai, E. Chuang and S.L. Zeichner, Dissection of the Kaposi's sarcoma-associated herpesvirus gene expression program by using the viral DNA replication inhibitor cidofovir, *Journal of Virology*, **78** (2004), 13637–13652.

[20] D.M. Lukac, R. Renne, J.R. Kirshner and D. Ganem, Reactivation of Kaposi's sarcoma-associated herpesvirus infection from latency by expression of the ORF50 transactivator, a homolog of the EBV R protein, *Virology*, **252** (1998), 304–312.

[21] R. Renne, M. Lagunoff, W.D Zhong and D. Ganem, The size and conformation of Kaposi's sarcoma-associated herpesvirus (human herpesvirus 8) DNA in infected cells and virions, *Journal of Virology*, **70** (1996), 8151–8154.

[22] G.A.F. Seber and C.J. Wild, C.J., *Nonlinear Regression*, John Wiley & Sons, Inc., NY, 1989.

[23] R. Sun, L. Su-Fang, K. Staskus, L. Gradoville, E. Grogan, A. Haase and G. Miller, Kinetics of Kaposi's sarcoma-associated herpesvirus gene expression, *Journal of Virology*, **73**(1999), 2232–2242.

[24] K. Takeuchi, Distribution of informational statistics and a criterion of model fitting, *Suri-Kagaku*(Mathematical Sciences), **153** (1976), 12–18.

[25] A. Tinari, P. Monini, M. Marchetti, M.G. Ammendolia, P. Leone, B. Ensoli and F. Superti, Lytic growth of human herpesvirus 8: Morphological aspects, *Ultrastructural Pathology*, **24** (2000), 301–310.

[26] Y. Yu, J.B. Black, C.S. Goldsmith, P.J. Browning, K. Bhalla and M.K. Offermann, Induction of human herpesvirus-8 DNA replication and transcription by butyrate and TPA in BCBL-1 cells, *Journal of General Virology*, **80** (1999), 83–90.

[27] J.P. Zoeteweij, S.T. Eyes, J.M. Orenstein, T. Kawamura, L.J. Wu, B. Chandran, B. Forghani, and A. Blauvelt, Identification and rapid quantification of early- and late-lytic human herpesvirus 8 infection in single cells by flow cytometric analysis: Characterization of antiherpesvirus agents, *Journal of Virology*, **73** (1999), 5894–5902.